SURVIVAL PREDICTION USING MACHINE LEARNING

The Model Predicts Whether A Passenger Would Survive On The Titanic Taking Into Account And Comparing And Finding Relations Amongst Various Features.

What is Survival Analysis?

Survival analysis is **a statistical method that aims to predict the time to an event, such as death, the diagnosis of a disease or the failure of a mechanical part**. A key aspect of survival analysis is the presence of censored data, indicating that the event of interest has not occurred during the study period

The objective in survival analysis — also referred to as reliability analysis in engineering — is to establish a connection between covariates and the time of an event. The name *survival analysis* originates from clinical research, where predicting the time to death, i.e., survival, is often the main objective. Survival analysis is a type of regression problem (one wants to predict a continuous value), but with a twist. It differs from traditional regression by the fact that parts of the training data can only be partially observed – they are *censored*.

For other survival models that do not rely on the proportional hazards assumption, it is often impossible to estimate survival or cumulative hazard function. Their predictions are risk scores of arbitrary scale. If samples are ordered according to their predicted risk score (in ascending order), one obtains the sequence of events, as predicted by the model. This is the return value of the **predict()** method of **all survival models in scikit-survival**.

Consequently, predictions are often evaluated by a measure of rank correlation between predicted risk scores and observed time points in the test data. In particular, Harrell's concordance index (**sksurv.metrics.concordance_index_censored()**) computes the ratio of correctly ordered (concordant) pairs to comparable pairs and is the default performance metric when calling a survival model's **score()** method.

Survival Data

As described in the section *What is Survival Analysis?* above, survival times are subject to right-censoring, therefore, we need to consider an individual's status in addition to survival time. To be fully compatible with scikit-learn, `Status` and `Survival_in_days` need to be stored as a structured array with the first field indicating whether the actual survival time was observed or if was censored, and the second field denoting the observed survival time, which corresponds to the time of death (if `Status == 'dead'`, δ=1) or the last time that person was contacted (if `Status == 'alive'`, δ=0).

Multivariate Survival Models

In the Kaplan-Meier approach used above, we estimated multiple survival curves by dividing the dataset into smaller sub-groups according to a variable. If we want to consider more than 1 or 2 variables, this approach quickly becomes infeasible, because subgroups will get very small. Instead, we can use a linear model, Cox's proportional hazard's model, to estimate the impact each variable has on survival.

First however, we need to convert the categorical variables in the data set into numeric values.

This article was published as a part of the Data Science Blogathon

Hey Folks, in this article, we will be understanding, how to analyze and predict, whether a person, who had boarded the RMS Titanic has a chance of survival or not, using Machine Learning's Logistic Regression model.

Brief description about Logistic Regression:

A simple yet crisp description of Logistic Description would be, "it is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes." as stated in the tutorial points article.

The graph of logistic regression is as shown below:

This article was published as a part of the Data Science Blogathon

Hey Folks, in this article, we will be understanding, how to analyze and predict, whether a person, who had boarded the RMS Titanic has a chance of survival or not, using Machine Learning's Logistic Regression model.

Brief description about Logistic Regression:

A simple yet crisp description of Logistic Description would be, “it is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.” as stated in the tutorial points article.

The graph of logistic regression is as shown below:

In This Blog-Post, We Would Be Going Through The Process Of Creating A Machine Learning Model Based On The Famous Titanic Dataset. This Gives The Titanic Survival Prediction, Taking Into Account Multiple Factors Such As- Economic Status (Class), Sex, Age, Etc.

The Model Predicts Whether A Passenger Would Survive On The Titanic Taking Into Account And Comparing And Finding Relations Amongst Various Features.

You Can Download The Official Titanic Dataset
From Https://Www.Kaggle.Com/C/Titanic/Data
We Start By Importing All The Important Packages /Libraries That Would Be Required For Building Our Model As Well As To Analyze The Given Datasets.

is the default performance metric when calling a survival model’s **score()** method.

Survival Data

As described in the section *What is Survival Analysis?* above, survival times are subject to right-censoring, therefore, we need to consider an individual’s status in addition to survival time. To be fully compatible with scikit-learn, `Status` and `Survival_in_days` need to be stored as a structured array with the first field indicating whether the actual survival time was observed or if was censored, and the second field denoting the observed survival time, which corresponds to the time of death (if `Status == 'dead'`, δ=1) or the last time that person was contacted (if `Status == 'alive'`, δ=0).

Multivariate Survival Models

In the Kaplan-Meier approach used above, we estimated multiple survival curves by dividing the dataset into smaller sub-groups according to a variable. If we want to consider more than 1 or 2 variables, this approach quickly becomes infeasible, because subgroups will get very small. Instead, we can use a linear model, Cox’s proportional hazard’s model, to estimate the impact each variable has on survival.

First however, we need to convert the categorical variables in the data set into numeric values.

This article was published as a part of the Data Science Blogathon

Hey Folks, in this article, we will be understanding, how to analyze and predict, whether a person, who had boarded the RMS Titanic has a chance of survival or not, using Machine Learning's Logistic Regression model.

Brief description about Logistic Regression:

A simple yet crisp description of Logistic Description would be, "it is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes." as stated in the tutorial points article.

The graph of logistic regression is as shown below:

As we now know what we have to do, to accomplish this task, we shall begin with the very first and the most important thing needed in machine learning, a **Dataset.**

What is a dataset:

A **data set,** as the name suggests, is a collection of data. In Machine Learning projects, we need a training **data set**. It is the actual **data set** used to train the model for performing various actions.

Here, in this case, we will be using a dataset available on the internet. One can find various such datasets over the internet.

The dataset that I've used in my code was the data available on Kaggle. You can also download it from here.

One thing must be kept in mind, the larger the data, the more we can train our model, and the more accurate our results come out to be. Don't worry if all of this sounds weird to you, it will all make sense in a few minutes.

ship to survive survived 1 of gender of patient is male and age 25?

Total samples are 891 or 40% of the actual number of passengers on board the Titanic (2,224). Survived is a categorical feature with 0 or 1 values. Around 38% samples survived representative of the actual survival rate at **32%**.

Workflow stages

The competition solution workflow goes through seven stages described in the Data Science Solutions book.

1. Question or problem definition.
2. Acquire training and testing data.
3. Wrangle, prepare, cleanse the data.
4. Analyze, identify patterns, and explore the data.
5. Model, predict and solve the problem.
6. Visualize, report, and present the problem solving steps and final solution.
7. Supply or submit the results.

The workflow indicates general sequence of how each stage may follow the other. However there are use cases with exceptions.

- We may combine mulitple workflow stages. We may analyze by visualizing data.
- Perform a stage earlier than indicated. We may analyze data before and after wrangling.
- Perform a stage multiple times in our workflow. Visualize stage may be used multiple times.
- Drop a stage altogether. We may not need supply stage to productize or service enable our dataset for a competition.

Why are Predictions Important in meachine learning?

- Machine learning model predictions **allow businesses to make highly accurate guesses as to the likely outcomes of a question based on historical data**, which can be about all kinds of things – customer .
- **Why prediction is important in machine learning?**
- churn likelihood, possible fraudulent activity, and more

Data Analysis And Visualization | Titanic Survival Prediction

The Next Step Is To Start Analyzing The Given Test And Train Datasets To Find Out Patterns Between The Features And Finding Relations Of Essential Features With The Target Feature (Survived Or Not).

We Observe That The Training Dataset Contains Approx 891 Rows And 12 Columns (Features) On The Other Hand The Testing Dataset Contains 418 Rows And 11 Columns (Since The Target Feature Of Survived Has Been Excluded For Us To Predict Analyzing The Training Dataset).

Now We Could Further Check For The Null Values Present In The Datasets.

You Would Find That In The Training Dataset The Age Column Contains 177 Null Values And The Embarked Column Contains About 2 Null Values Whereas For The Test Dataset The Age Column Contains 86 Null Values And The Cabin Column Contains About 327 Null Values. The Rest Of The Columns Have Values Properly Filled (We Would Handle The Null Values Later On, For Now Just Analyze And Note Them Down).

1. Survival Based On Passenger Class (P-Class)

2. Survival Based On Age

3. Survival Based Upon Embarked Label

Now We Could Even Find The Survival Rate Dependency Comparing With Individual Labels Without Plotting Them Like Above:-

1. Considering Passenger Class (P-Class)
2. Considering The SibSp Label
3. Considering The Embarked Label

eature Selection | Titanic Survival Prediction:-

Now We Saw That There Were Approx 12 Different Feature Columns Provided In The Dataset. Now Not All Features Have An Impact On The Required Target Feature (Survival In Our Case). So It Would Be Better To Select The Features Of

Major Importance And Drop Certain Features That Have A Minor Impact On Our Target Column. This Process Is Referred To As Feature Selection In Machine Learning.

Model Building And Training:-

Firstly Import The Train_test_split From The Sklearn Library And Then Split The Dataset Into Train And Test Specifying A Test_size.

Now Comes The Part Of Selecting An Algorithm For Training Our Model. I Prefer Taking Various Algorithms And Comparing Their Accuracy Score And Then Selecting The Best Fit Model Having The Maximum Score.

1. Logistic Regression Mode

2. SVM Model

3. KNN Model

4. Gaussian Naive_bayes Model

5. Decision Tree Model

Conclusion

(The Scores Obtained Can Be Further Improved By Hyperparameter Tuning. As They Say, There's Always A Massive Scope For Improvement Just Keep On Experimenting Till You Build Your Perfect Model … !).

Hope This Blog Helped You To Understand The Basic Concepts Of Building A Model For Solving The Titanic Survival Prediction Problem Statement.

S180850
S180311